Privacy in Big Data

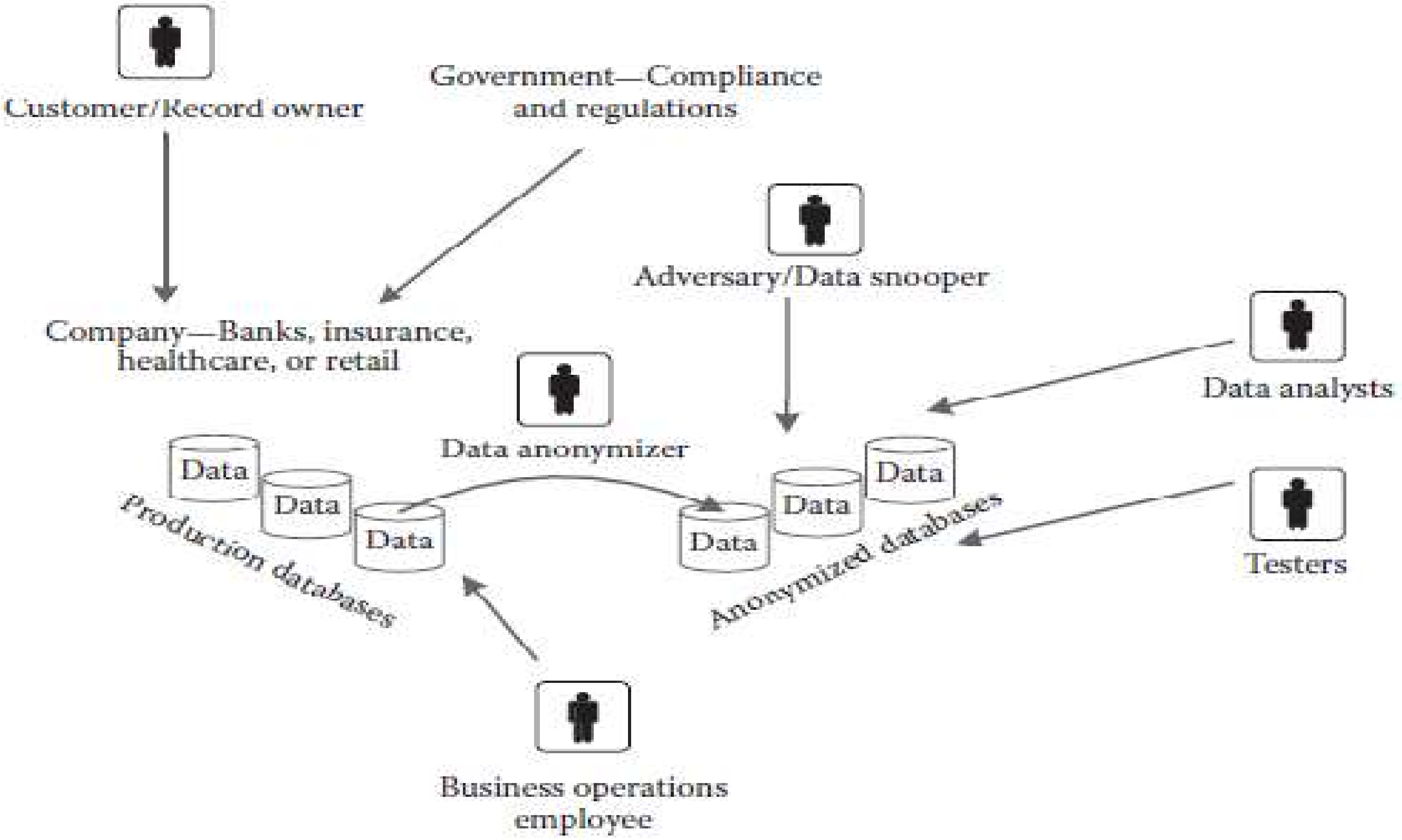
What Is Data Privacy?

- Thousands of ambulance service staff and housing benefits claimants have had their personal information accidently leaked in the latest UK data breach blunder (January 4, 2014; news in www.infosecurity-magazine.com/news/thousands-of-personal/-details.)
- Natural healthcare chain Community Health Systems (CHS) says that about 4.5 million pieces of “non-medical patient identification data related to our physician practice” have been stolen …. August 18, 2014/News www.infosecurity-magazine.com/news/45-million-records-stolen-from/
- NASDAQ-listed outsourcing firm EXL Service has lost a key client due to the breach of confidential client data by some of its employees.

What Is Data Privacy?

- There are numerous such incidents where customers' confidential personal information has been attacked by or lost to a data snooper.
- When such untoward incidents occur, organizations face legal suits, financial loss, loss of image, and, importantly, the loss of their customers.
- There are many stakeholders of data privacy in an organization

What Is Data Privacy?



What Is Data Privacy?

- There are many stakeholders of data privacy in an organization
- Company:
  - Any organization like a bank, an insurance company, or an e-commerce, retail, healthcare, or social networking company that holds large amounts of customer-specific data.
  - They are the custodians of customer data, which are considered very sensitive, and have the responsibility of protecting the data at all costs.
  - Any loss of these sensitive data will result in the company facing legal suits, financial penalties, and loss of reputation.

What Is Data Privacy?

- Customer/record owner:
  - An organization's customer could be an individual or another organization who share their data with the company.
  - For example, an individual shares his personal information, also known as PII, such as his name, address, gender, date of birth, phone numbers, e-mail address, and income with a bank.
  - PII is considered sensitive as any disclosure or loss could lead to undesired identification of the customer or record owner.
  - It has been shown that gender, age and zip code are sufficient to identify a large population of people in the United States.

What Is Data Privacy?

Government:

- Government defines what data protection regulations that
- the company should comply with.
- Examples of such regulations are the HIPPA Act, the EU Data Protection Act, and the Swiss Data Protection Act.
- It is mandatory for companies to follow government regulations on data protection..

Data anonymizer:

- A person who anonymizes and provides data for analysis or as test data.

What Is Data Privacy?

Data analyst:

- This person uses the anonymized data to carry out data mining activities like prediction, knowledge discovery, and so on.
- Following government regulations, such as the Data Moratorium Act, only anonymized data can be used for data mining.
- Therefore, it is important that the provisioned data support data mining functionalities.

What Is Data Privacy?

Tester:

- Outsourcing of software testing is common among many companies.
- High-quality testing requires high-quality test data, which is present in production systems and contains customer-sensitive information.
- In order to test the software system, the tester needs data to be extracted from production systems, anonymized, and provisioned for testing.
- Since test data contain customer-sensitive data, it is mandatory to adhere to regulatory compliance in that region/country.

What Is Data Privacy?

Business operations employee:

- Data analysts and software testers use anonymized data that are at rest or static, whereas business operations employees access production data because they need to support customer’s business requirements.
- Business operations are generally outsourced to BPO (business process outsourcing) companies.
- In this case too, there is a requirement to protect customer-sensitive data but as this operation is carried out during run-time, a different set of data protection techniques are required to protect data from business operations employees.

What Is Data Privacy?

Adversary/data snooper:

- Data are precious and their theft is very common.
- An adversary can be internal or external to the organization.
- The anonymization design should be such that it can thwart an adversary’s effort to identify a record owner in the database.

Protecting Sensitive Data

- “I know where you were yesterday!” Google knows your location when you use Google Maps.
- Google maps can track you wherever you go when you use it on a smart phone.
- Mobile companies know your exact location when you use a mobile phone.
- You have no place to hide. You have lost your privacy.
- This is the flip side of using devices like smart phones, Global positioning systems (GPS), and radio frequency identification (RFID).
- Why should others know where you were yesterday? Similarly, why should others know your health issues or financial status?
- All these are sensitive data and should be well protected as they could fall into the wrong hands and be exploited.

Protecting Sensitive Data

Data D in the tables contains four disjointed data sets:

1. *Explicit identifiers* (*EI*): Attributes that identify a customer (also called record owner) directly. These include attributes like social security number (SSN), insurance ID, and name.

2. *Quasi-identifiers* (*QI*): Attributes that include geographic and demographic information, phone numbers, and e-mail IDs. Quasiidentifiers are also defined as those attributes that are publicly available, for example, a voters database.

3. *Sensitive data* (*SD*): Attributes that contain confidential information about the record owner, such as health issues, financial status, and salary, which cannot be compromised at any cost.

4. *Nonsensitive data* (*NSD*): Data that are not sensitive for the given context.

Protecting Sensitive Data

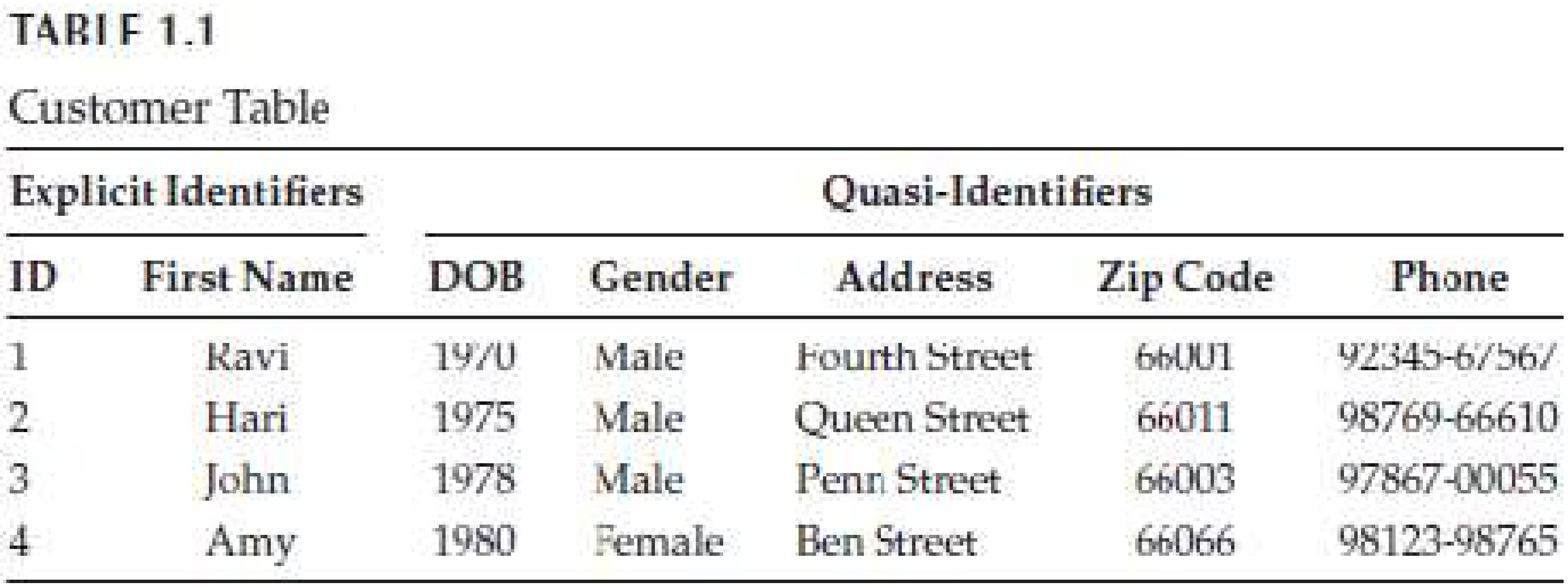


TABLE 1.1

Customer Table

Explicit Identifiers		Quasi-Identifiers				
ID	First Name	DOB	Gender	Address	Zip Code	Phone
1	Ravi	1970	Male	Fourth Street	66001	92345-67567
2	Hari	1975	Male	Queen Street	66011	98769-66610
3	John	1978	Male	Penn Street	66003	97867-00055
4	Amy	1980	Female	Ben Street	66066	98123-98765

Protecting Sensitive Data

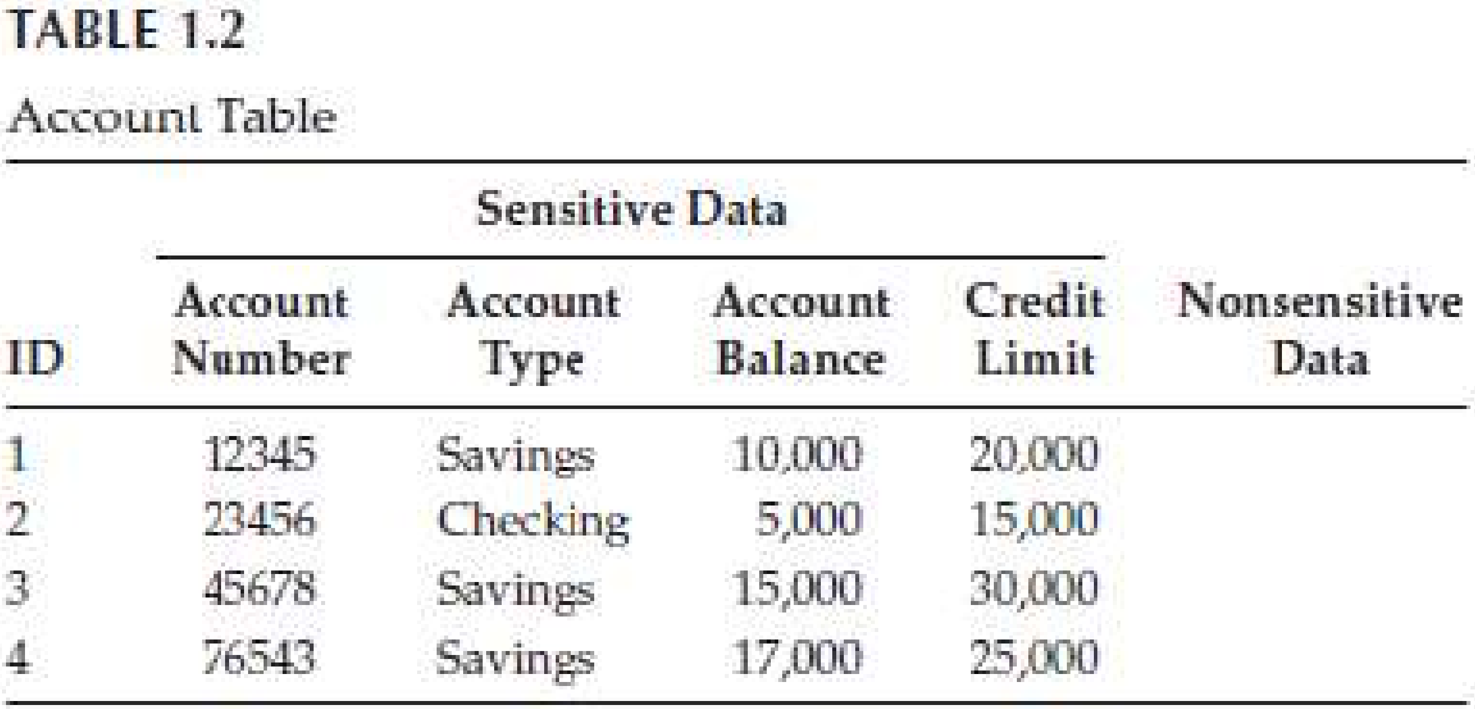


TABLE 1.2

Account Table

	Sensitive Data				
ID	**Account Number**	**Account Type**	**Account Balance**	**Credit Limit**	**Nonsensitive Data**
1	12345	Savings	10,000	20,000	
2	23456	Checking	5,000	15,000	
3	45678	Savings	15,000	30,000	
4	76543	Savings	17,000	25,000	

Protecting Sensitive Data

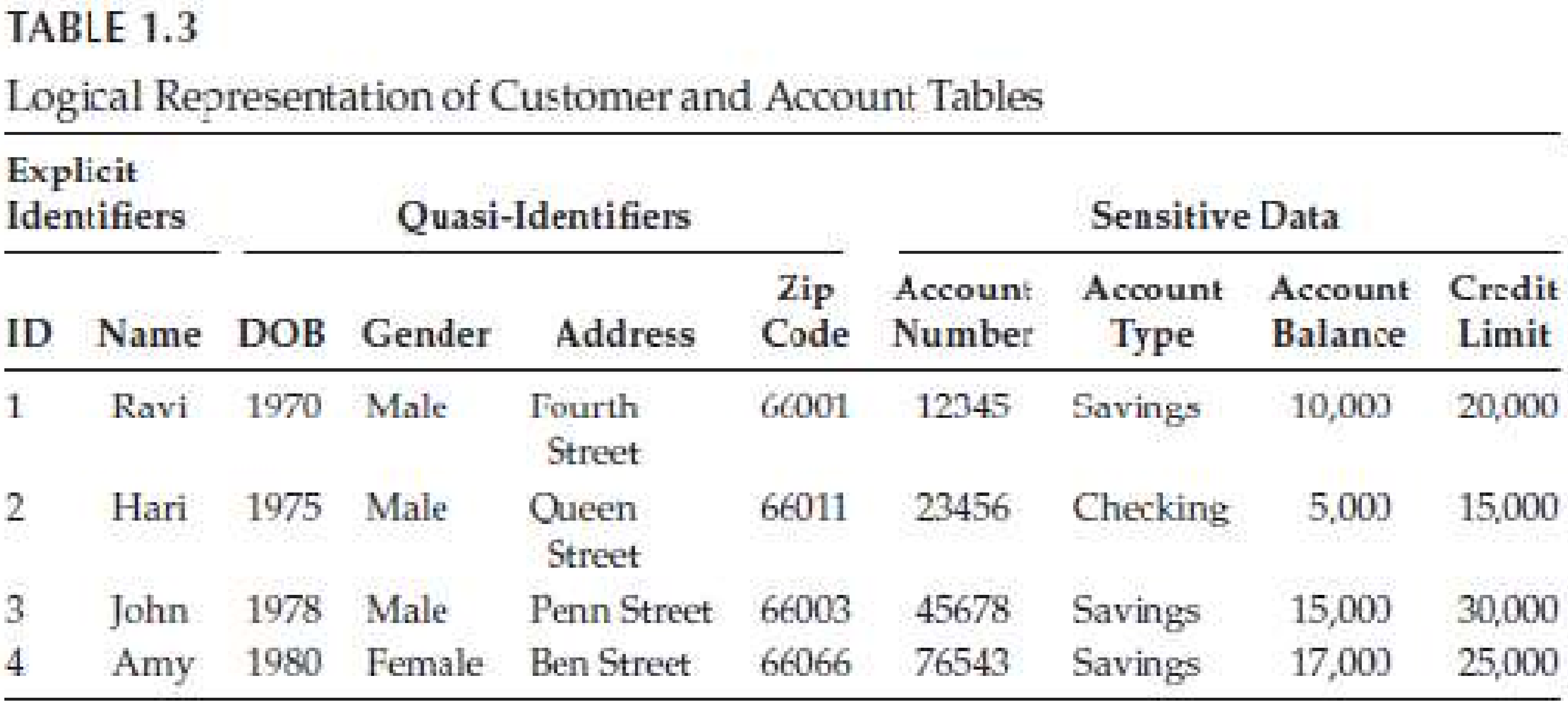


TABLE 1.3

Logical Representation of Customer and Account Tables

Explicit Identifiers		Quasi-Identifiers				Sensitive Data			
ID	Name	DOB	Gender	Address	Zip Code	Account Number	Account Type	Account Balance	Credit Limit
1	Ravi	1970	Male	Fourth Street	66001	12345	Savings	10,000	20,000
2	Hari	1975	Male	Queen Street	66011	23456	Checking	5,000	15,000
3	John	1978	Male	Penn Street	66003	45678	Savings	15,000	30,000
4	Amy	1980	Female	Ben Street	66066	76543	Savings	17,000	25,000

Protecting Sensitive Data

- The first two data sets, the EI and QI, uniquely identify a record owner and when combined with sensitive data become sensitive or confidential.
- The data set D is considered as a matrix of m rows and n columns.
- Matrix D is a vector space where each row and column is a vector

$$D = [D_{EI}]\ [D_{QI}]\ [D_{SD}] \qquad (1.1)$$

- Each of the data sets, EI, QI, and SD, are matrices with m rows and i, j, and k columns, respectively.
- We need to keep an eye on the index j (representing QI), which plays a major role in keeping the data confidential.

Protecting Sensitive Data

- ➢ Apart from assuring their customers' privacy, organizations also have to comply with various regulations in that region/country, as mentioned earlier.
- ➢ Most countries have strong privacy laws to protect citizens' personal data.
- ➢ Organizations that fail to protect the privacy of their customers or do not comply with the regulations face stiff financial penalties, loss of reputation, loss of customers, and legal issues.
- ➢ This is the primary reason organizations pay so much attention to data privacy
- ➢ data protection techniques, such as cryptography and anonymization, are used prior to sharing data.

Privacy and Anonymity

- Anonymization is a process of logically separating the identifying information (PII) from sensitive data.
- Referring to Table 1.3, the anonymization approach ensures that EI and QI are logically separated from SD.
- As a result, an adversary will not be able to easily identify the record owner from his sensitive data.
- privacy and anonymity are flip sides of the same coin

Privacy and Anonymity

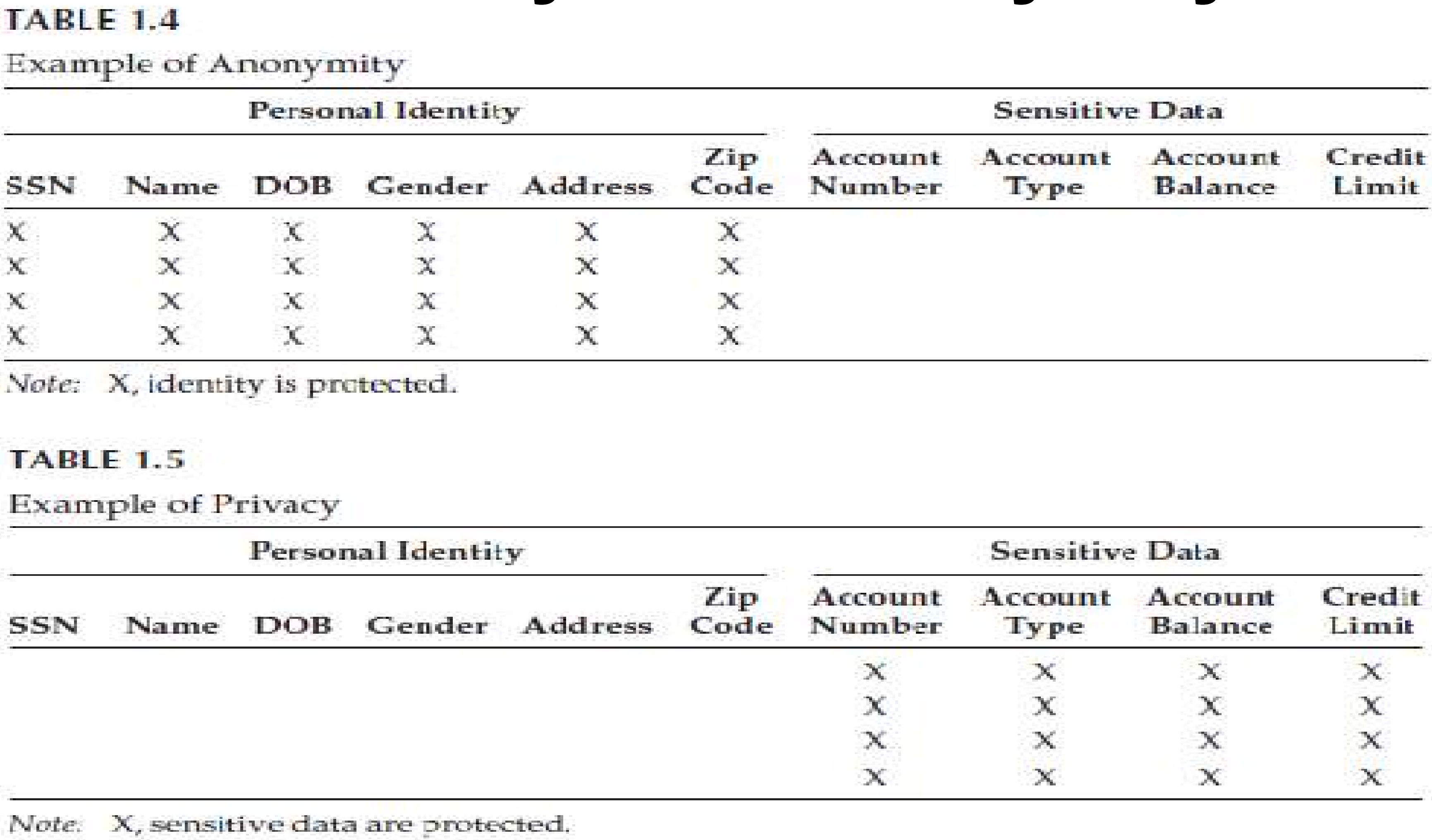


TABLE 1.4

Example of Anonymity

Personal Identity						Sensitive Data			
SSN	Name	DOB	Gender	Address	Zip Code	Account Number	Account Type	Account Balance	Credit Limit
X	X	X	X	X	X				
X	X	X	X	X	X				
X	X	X	X	X	X				
X	X	X	X	X	X				

Note: X, identity is protected.

TABLE 1.5

Example of Privacy

Personal Identity						Sensitive Data			
SSN	Name	DOB	Gender	Address	Zip Code	Account Number	Account Type	Account Balance	Credit Limit
						X	X	X	X
						X	X	X	X
						X	X	X	X
						X	X	X	X

Note: X, sensitive data are protected.

Privacy and Anonymity

- There is a subtle difference between privacy and anonymity.
- The word privacy is also used in a generic way to mean anonymity, and there are specific use cases for both of them.
- Table 1.4 illustrates an anonymized table where PII is protected and sensitive data are left in their original form.
- Sensitive data should be in original form so that the data can be used to mine useful knowledge.
- Anonymization is a two-step process: data masking and de-identification.

Privacy and Anonymity

- Data masking is a technique applied to systematically substitute, suppress, or scramble data that call out an individual, such as names, IDs, account numbers, SSNs, etc.
- Masking techniques are simple techniques that perturb original data.
- ➢ De-identification is applied on QI fields. QI fields such as date of birth, gender, and zip code have the capacity to uniquely identify individuals.
- ➢ Combine that with SD, such as income, and a Warren Buffet or Bill Gates is easily identified in the data set.
- ➢ By de-identifying, the values of QI are modified carefully so that the relationship is till maintained by identities cannot be inferred.

Privacy and Anonymity

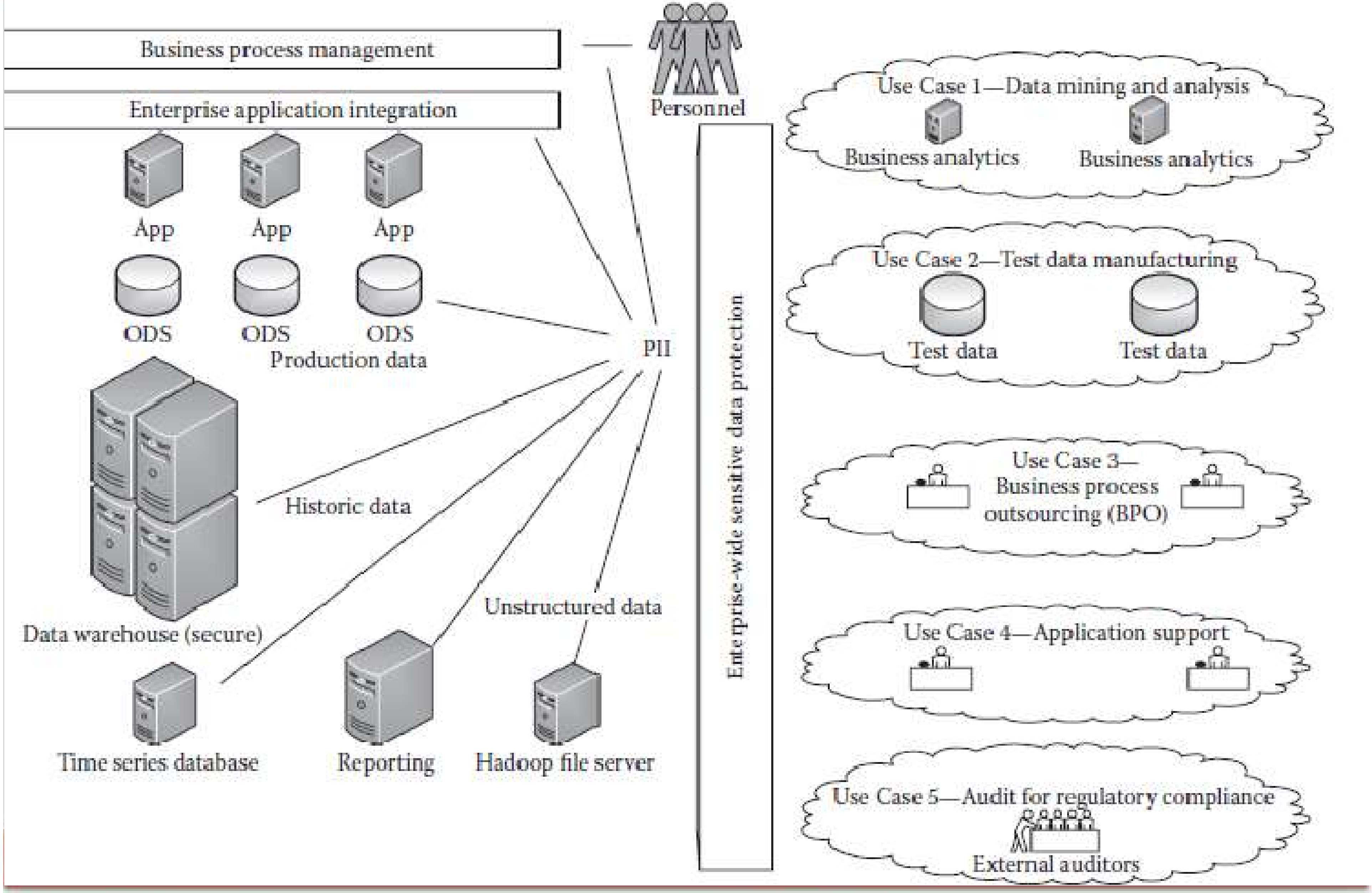
- the original data set is D which is anonymized, resulting in data set

$$D' = T(D) \text{ or } T([D_{EI}][D_{QI}][D_{SD}])$$

- where T is the transformation function.
- As a first step in the anonymization process, EI is completely masked and no longer relevant in D’.
- As mentioned earlier, no transformation is applied on SD and it is left in its original form.
- This results in D’ = T($[D_{QI}]$), which means that transformation is applied only on QI as EI is masked and not considered as part of D’ and SD is left in its original form.
- D’ can be shared as QI is transformed and SD is in its original form but it is very difficult to identify the record owner.

Need for Sharing Data

- Organizations tend to share customer data as there is much insight to be gained from customer-sensitive data.
- For example, a healthcare provider’s database could contain how patients have reacted to a particular drug or treatment.
- This information would be useful to a pharmaceutical company.
- However, these sensitive data cannot be shared or released due to legal, financial, compliance, and moral issues.
- But for the benefit of the organization and the customer, there is a need to share these data responsibly, which means the data are shared without revealing the PII of the customer.


Business process management
Enterprise application integration
Personnel
App
App
App
ODS
ODS
ODS
Production data
PII
Data warehouse (secure)
Historic data
Unstructured data
Time series database
Reporting
Hadoop file server
Enterprise-wide sensitive data protection
Use Case 1—Data mining and analysis
Business analytics
Business analytics
Use Case 2—Test data manufacturing
Test data
Test data
Use Case 3—
Business process
outsourcing (BPO)
Use Case 4—Application support
Use Case 5—Audit for regulatory compliance
External auditors

Need for Sharing Data

Let us now explore for what purposes the data are shared and how.

- Data mining and analysis
- Application testing
- Business operation
- Application support
- Auditing and reporting for regulatory compliance

These use cases can be classified under two categories:

1. Privacy protection of sensitive data at rest
2. Privacy protection of sensitive data in motion (at run-time)

Methods of Protecting Data

- Cryptographic techniques are probably one of the oldest known techniques for data protection
- Anonymization is a set of techniques used to modify the original data in such a manner that it does not resemble the original value but maintains the semantics and syntax
- Tokenization is a data protection technique that has been extensively used in the credit card industry but is currently being adopted in other domains as well. Tokenization is a technique that replaces the original sensitive data with nonsensitive placeholders referred to as tokens

Balancing Data Privacy and Utility

- Privacy preservation should also ensure utility of data. In other words, the provisioned data should protect the individual's privacy and at the same time ensure that the anonymized data are useful for knowledge discovery.
- By anonymizing the data, EI are completely masked out, QI is de-identified by applying a transformation function, and SD is left in its original form.
- There is a strong correlation between QI and SD fields. So, as part of privacy preservation, this correlation between QI fields and SD fields should not be lost.
- If the correlation is lost, then the resulting data set is not useful for any purpose.
- As a transformation function is applied on QI, it is obvious that the correlation between QI fields and SD fields is affected or weakened, and this indicates how useful the transformed data are for the given purpose